

Abstract

An organization's data records are often noisy: because of transcription errors, incomplete information, and lack of standard formats for textual data. A fundamental task during data cleansing and integration is matching strings -perhaps across multiple relations- that refer to the same entity (e.g., organization name or address). Furthermore, it is desirable to perform this matching within an RDBMS, which is where the data is likely to reside. In this paper, We adapt the widely used and established cosine similarity metric from the information retrieval field to the relational database context in order to identify potential string matches across relations. We then use this similarity metric to characterize this key aspect of data cleansing and integration as a join between relations on textual attributes, where the similarity of matches exceeds a specified threshold. Computing an exact answer to the text join can be expensive. For query processing efficiency, we propose an approximate, sampling-based approach to the join problem that can be easily and efficiently executed in a standard, unmodified RDBMS. Therefore the present invention includes a system for string matching across multiple relations in a relational database management system comprising generating a set of strings from a set of characters, decomposing each string into a subset of tokens, establishing at least two relations within the strings, establishing a similarity threshold for the relations, sampling the at least two relations, correlating the relations for the similarity threshold and returning all of the tokens which meet the criteria of the similarity threshold.